Antonis Maronikolakis, PhD Student

github email linkedin website scholar

Education

PhD, Supervised by Hinrich Schütze — January 2020 - Current
Ludwig-Maximilians-Universität, Munich, Germany
Topic: Multilingual Data and Model Development for Social Media Text

- Worked on **Social NLP**, curating parody and hate speech datasets, as well as analyzing existing datasets for bias. Proposed mitigation strategies for gender and racial biases in multilingual settings.
- Examined **few-shot learning** and **multilinguality** capabilities of large language models.
- **Worked on a grant proposal**, receiving funding from the ERC for our proof-of-concept project Respond2Hate.
- **Supervised student theses**, publishing two papers on hate speech detection and bias analysis, as well as two papers in the field of generated text analysis and domain adaptation.

MSc Speech and Language Processing - First Honours (1:1) — September 2018 - September 2019
University of Sheffield, Sheffield, UK
Thesis: Text Generation for News Headlines

BSc Computer Science - 8.4/10.0 (Upper 2:1) — October 2014 - June 2018
University of Piraeus, Athens, Greece
Thesis: Approximation and Heuristic Algorithms for Ridesharing

Professional Experience

Junior Applied Scientist at Zalando
Customer Reviews Moderation and Ranking — June 2022 - November 2022

- Worked in the domain of fashion customer reviews, researching methods to improve reviews moderation and ranking.
- Developed a few-shot learning model to improve reviews moderation in a multilingual setting.
  - Finetuned BERT models using pattern-exploiting few-shot learning (PET).
  - Showed that we can improve performance in lower-resource settings with multilingual transfer.
  - Run a Monte Carlo simulation to identify training examples most conducive to performance.
  - Our method outperforms production models using a fraction of the training data (32 vs. 20k).
- Explored methods to improve ranking of customer reviews, aiming to rank higher-information reviews closer to the top for better visibility.
  - Oversaw an A/B test to compare our proposed ranking algorithm with the current method.
- Researched methods to automatically identify language of reviews, speeding up review moderation and translation efforts.
- Undertook onboarding duties for newer researchers when my lead had to take an extended leave.

NLP Researcher at AI4Dignity
Hate Speech Data Analysis for Marginalized Global Communities — January 2021 - December 2021

- Worked in a multi-disciplinary team of NLP researchers, anthropologists and fact-checkers.
- Led the data collection and annotation team, culminating in a dataset of 20k examples spanning 4 countries (Brazil, Germany, India and Kenya).
- Developed machine learning models, performed data analysis and deployed a proof-of-concept tool to combat hate speech in the examined countries.
- Our work has been published at Findings of ACL 2022. We have also published a policy brief at the EU Commission and a paper on ethical scaling of hate speech detection in a sociology journal.

Speech Processing Practitioner for VoiceBase Research Lab
Analyzed Variational Autoencoders and Phrase Detection methods — November 2018 - August 2019

- Joined Dr. Thomas Hain's research lab as a student research assistant.
- Assisted researchers on work with Variational Autoencoders, Phrase Detection and Extraction.

AI Programmer/Writer for Google Summer of Code
Mentored under Dr. Peter Norvig, working on AI algorithms — June 2017 - September 2017

- Worked on the Python repository of *Artificial Intelligence: A Modern Approach*, implementing and writing about NLP and ML algorithms, as well as polishing pseudocode for the book.

Publications

2023

Sociocultural knowledge is needed for selection of shots in hate speech detection tasks. Antonis Maronikolakis, Abdullatif Köksal, Hinrich Schütze. Pre-print.

This joke is [MASK]: Recognizing Humor and Offense with Prompting. Junze Li, Mengjie Zhao, Yubo Xie, Antonis Maronikolakis, Pearl Pu, Hinrich Schütze. Transfer Learning for Natural Language Processing Workshop (NeurIPS).

2022

Listening to Affected Communities to Define Extreme Speech: Dataset and Experiments. Antonis Maronikolakis, Axel Wisiorek, Leah Nann, Haris Jabbar, Sahana Udupa and Hinrich Schütze. 2022. In Findings of the Association for Computational Linguistics: ACL 2022.

Ethical Scaling for Content Moderation: Extreme Speech and the (In)Significance of Artificial Intelligence. Sahana Udupa, Antonis Maronikolakis, Hinrich Schütze, Axel Wisiorek. 2022. Harvard Kennedy School, Shorenstein Center.

Separating Hate Speech and Offensive Language Classes via Adversarial Debiasing Shuzhou Yuan, Antonis Maronikolakis, Hinrich Schütze. 2022. The 6th Workshop on Online Abuse and Harms (NAACL).

Analyzing Hate Speech Data along Racial, Gender and Intersectional Axes. Antonis Maronikolakis, Philip Baader, Hinrich Schütze. 2022. 4th Workshop on Gender Bias in Natural Language Processing (NAACL).

2021

Wine is Not v i n. – On the Compatibility of Tokenizations Across Languages. Antonis Maronikolakis, Philipp Dufter, and Hinrich Schütze. 2021. In Findings of the Association for Computational Linguistics: EMNLP 2021.

Artificial Intelligence, Extreme Speech, and the Challenges of Online Content Moderation. Sahana Udupa, Elonnai Hickok, Antonis Maronikolakis, Hinrich Schuetze, Laura Csuka, Axel Wisiorek, Leah Nann. 2021. AI4Dignity Project. EU Commission Policy Brief.

BERT Cannot Align Characters. Antonis Maronikolakis, Philipp Dufter, and Hinrich Schütze. 2021. In Proceedings of the Second Workshop on Insights from Negative Results in NLP (EMNLP).

Identifying Automatically Generated Headlines Using Transformers. Antonis Maronikolakis, Hinrich Schütze, and Mark Stevenson. 2021. In Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda (NAACL).

Multidomain Pretrained Language Models for Green NLP. Antonis Maronikolakis and Hinrich Schütze. 2021. In Proceedings of the Second Workshop on Domain Adaptation for NLP (EACL).

2020

Analyzing Political Parody in Social Media. Antonis Maronikolakis, Danae Sánchez Villegas, Daniel Preotiuc-Pietro, and Nikolaos Aletras. 2020. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

Talks & Awards

- New approaches to dealing with content moderation challenges - ActiveFence - 2022
- Introduction to AI4Dignity - Konvens 2021, MMHS Workshop - 2021
- Extreme Speech Detection with NLP - LMU, AI4Dignity Counterathon - 2021
- Scholarship for academic excellence - 2016-2017, 2017-2018
- Kaggle Machine Learning Competitions (2018) - Silver and Bronze Medals
- Microsoft Imagine Cup Competition (2015 and 2016) - National Finalist
- National Computer Science Competition 2014 - Finalist